Rethinking the Backbone in Model-Inversion-Based Exemplar-Free Continual Learning

Anonymous ECCV 2026 Submission

Paper ID #31

Abstract. In continual learning, inversion-based replay reconstructs surrogate inputs from the trained model and replays them instead of real past data, mitigating catastrophic forgetting in the exemplar-free setting. However, the success of such methods largely depends on signals that are difficult to extract from the model, as discriminative backbones discard a significant amount of information. Based on this observation, we study whether a backbone designed to be inverted could be used as the main component of such methods. To this end, we consider a pseudo-invertible backbone and compare it with a parameter-matched ResNet-32 in an online class-incremental learning setup. Our results on CIFAR-100 show that, among the data-free approaches we consider, inversion replay best preserves old-class knowledge, and that the pseudo-invertible backbone improves old-class accuracy from 9.5 to 18.6 under the same inversion setup.

Keywords: Continual learning · Exemplar-free · Model inversion · Invertible backbone · Catastrophic forgetting

1 Introduction

Continual learning (CL) updates a model on a sequence of tasks; we focus on class-incremental learning (CIL), its most challenging setting, where new classes are added while previous training data are no longer available, making deep models prone to catastrophic forgetting (CF) [6, 17, 20]. To mitigate this phenomenon, exemplar replay is a simple but strong mechanism to preserve old-task information [12,13]. However, in exemplar-free settings that rule out storing real samples, the challenge is to recover a useful replay signal without a buffer.

Besides regularization and parameter isolation [20], exemplar-free CL can *replay* a signal recovered from the model; such methods differ mainly in where it is produced. Generative replay synthesizes samples with a separate generator [4, 14, 18], trained on the fly or pretrained [5], while latent or prototype replay stores and reuses compact representations rather than images [9,10]. Model-inversion approaches instead synthesize pseudo-inputs around a trained model, using constraints on features, batch-normalization statistics, or class-wise distributions [11, 15, 16, 19]. These methods improve how replay is generated, but the backbone is rarely treated as the main design variable.

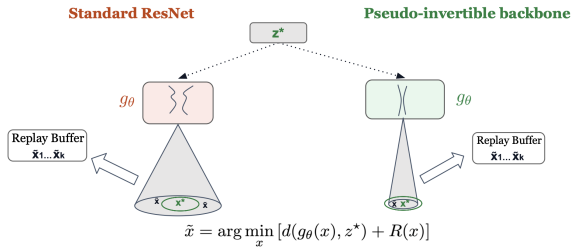


Fig. 1: Backbone-aware model inversion. Each cone is the set of inputs matching one old-class target: wide and ambiguous for a ResNet-32, contracted to a substantially less ambiguous preimage by the pseudo-invertible backbone, giving usable replay and better retention.

Discriminative backbones are hard to invert: pooling discards spatial detail and non-injective activations collapse distinct inputs to one response, leaving inversion underconstrained and reliant on extra statistics or losses to be usable for replay [11, 16, 19]. The difficulty thus stems partly from the backbone, not only the inversion procedure.

We therefore treat the backbone, not the inversion objective, as the design variable for model-inversion replay (Fig. 1). i-ResNets give a fully invertible reference point, since their contractive residual blocks admit a tractable inverse [1]; we relax this into a *pseudo*-invertible backbone that keeps a practical inverse path even under online single-pass training. On online CIFAR-100 this yields better old-class retention than a same-budget non-invertible ResNet under identical inversion, against no-replay and real-old-oracle references.

2 Backbone-Aware Model Inversion

Let $g_\theta : \mathcal{X} \to \mathcal{Z}$ be a classifier backbone and let $z^\star$ be an internal replay target. Standard model inversion retrieves a pseudo-input $\hat{x}$ by solving

$$\hat{x} = \arg\min_x d(g_\theta(x), z^\star) + R(x), \tag{1}$$

where R encodes data priors or external constraints, and d is a metric in the latent space $\mathcal{Z}$. This optimization is often underconstrained. Consequently, the less information the backbone preserves, the more the inversion depends on R to produce usable pseudo-inputs. This is further exacerbated in online CL: a model trained for a single online pass yields much noisier inversions than one trained for 20 offline epochs (Fig. 2), pointing to lower feature recoverability.

We instead reduce this ambiguity architecturally, building a *pseudo*-invertible backbone (the *morph*) from invertibility-preserving components. Following the invertible-network literature [1], it avoids destructive downsampling (an injective-padding stem and squeeze downsampling preserve dimensions) and keeps each residual block close to contractive through spectral normalization and a small

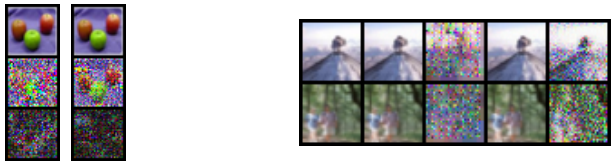


Fig. 2: Optimization-based inversion under matched settings. Left: a latent inverted through a ResNet-32 trained for a single online pass (left) vs. 20 offline epochs (right); far noisier online. Right: pseudo-invertible (morph) versus non-invertible (ResNet-32) backbones on two held-out examples; columns: original, untrained morph, untrained ResNet-32, trained morph, and trained ResNet-32.

per-stage residual scale, which makes the feature map near-injective and its inversion well-posed. Unlike strictly invertible networks, it uses standard batch normalization for discriminative training, inverted at replay time through its running statistics. Because we only encourage contraction rather than certify a Lipschitz bound, replay does not rely on a guaranteed fixed-point inverse: we recover the pseudo-input by minimizing the feature-matching objective of Eq. 1. Under matched settings, the morph reconstructions stay closer to the target than the ResNet-32's (Fig. 2).

3 Evaluation Protocol

The experiments are conducted on CIFAR-100 [7], training the classifier incrementally, online, on five tasks of 20 classes each, each seen in a single pass. Evaluation is performed without task identity, on all seen classes. We consider the exemplar-free class-incremental setting: no past images are stored (only compact per-class statistics), and old-class replay is synthesized by inverting the current model. The inversion recipe follows PMI [16], originally offline, adapted to this online regime where old-class information cannot be refined over many passes but only tracked as the model drifts. We compare three conditions: no replay, inversion replay, and real-old replay as an upper-bound oracle. ResNet-32 and the pseudo-invertible backbone are evaluated under the same replay protocol and at matched parameter count. A standard ResNet-32 ($\sim$0.47 M) is much smaller, so we width-double it ($\sim$1.86 M) to match the morph ($\sim$1.7 M); if anything it is slightly larger, so the gap is not a capacity effect. We report the accuracy over all seen classes measured after the final task (Acc), the accuracy on old (previously seen) classes (Old), and the forgetting measure (F), i.e. the mean per-task drop from peak to final accuracy (lower is better) [3].

4 Results and Discussion

(1) Retention: model inversion vs. regularization (Table 1). On the morph backbone, data-free regularization does not mitigate forgetting: LwF [8] and ACE [2]

Table 1: Online CIFAR-100 5×20, one pass per task. Top block: data-free methods on the morph backbone against the real-replay ceiling; only inversion replay (ours) recovers old-class accuracy. Bottom block: the same inversion recipe on a ResNet-32, whose oldest classes collapse. All methods are data-free except the real-old oracles (real images, upper bound). Inversion (ours) and the oracles are means over three seeds; the regularization baselines are shown at their most favorable seed, which favors them. T1/T2 are the two oldest tasks' final accuracy. MMD is the unbiased feature-space MMD (diagonal excluded) between inversions and real old-class features in each backbone's penultimate space, three-seed mean (std 0.07 and 0.18); lower is closer, n/a for real replay.

Method	Acc↑	Old↑	F↓	T1	T2	MMD↓
Morph backbone						
Online fine-tuning	7.8	0.0	38.8	–	–	–
LwF	8.4	0.1	36.3	–	–	–
ACE	9.8	4.2	27.4	–	–	–
Inversion (ours)	**17.6**	**18.6**	**6.3**	**17.8**	**14.4**	**0.17**
Real-old (oracle)	26.5	26.6	4.8	27.2	21.8	–
ResNet-32 backbone						
Inversion	12.1	9.5	13.5	1.6	1.5	0.44
Real-old (oracle)	27.3	26.3	6.2	26.8	21.2	–

remain at the fine-tuning floor, with little old-task retention (old ≤ 5, $F \geq 27$). Model-inversion replay, in contrast, retains old-class information (old 18.6, F 6.3), improving old-class retention from at most 4.2 (ACE) to 18.6 and recovering about 70% of the real-image oracle.

(2) The gain from the pseudo-invertible backbone. The same inversion-replay recipe fails on a ResNet-32: its oldest tasks collapse to near chance (T1 1.6, T2 1.5) while the morph keeps them alive (T1 17.8, T2 14.4). Under real replay the two backbones are on par (the ResNet-32 is if anything slightly stronger, 27.3 vs. 26.5 Acc), so the gap is not a capacity or backbone-strength effect. Under *inversion*, however, it is decisive: the morph reaches 17.6/18.6 (Acc/Old) against 12.1/9.5 for the ResNet, and on the oldest task it recovers about 65% of its real-replay oracle, the ResNet only 6%. This points to coverage: the ResNet's final-checkpoint inversions lie about 2.5× farther from the real old-class distribution than the morph's (feature-space MMD 0.44 vs. 0.17, Table 1), consistent with poorer coverage of that distribution rather than merely weaker optimization.

Conclusion. In online exemplar-free CL the bottleneck is old-class retention, which data-free regularization such as LwF and ACE leaves largely unaddressed. Model-inversion replay addresses it when the backbone is sufficiently amenable to inversion: our pseudo-invertible backbone recovers a significant part of the real-replay ceiling, where a parameter-matched ResNet-32 collapses. This positions backbone design as a promising direction for inversion-based continual learning, and a starting point for broader evaluation.

References

1. Behrmann, J., Grathwohl, W., Chen, R.T.Q., Duvenaud, D., Jacobsen, J.H.: Invertible residual networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 573–582. PMLR (2019), https://proceedings.mlr.press/v97/behrmann19a.html 2
2. Caccia, L., Aljundi, R., Asadi, N., Tuytelaars, T., Pineau, J., Belilovsky, E.: New insights on reducing abrupt representation change in online continual learning. In: ICLR (2022) 3
3. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.S.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: ECCV (2018) 3
4. Gao, R., Liu, W.: Ddgr: Continual learning with deep diffusion-based generative replay. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 10744–10763 (2023) 1
5. Kim, J., Cho, H., Kim, J., Tiruneh, Y.Y., Baek, S.: Sddgr: Stable diffusion-based deep generative replay for class incremental object detection (2024), https://arxiv.org/abs/2402.17323 1
6. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (Mar 2017). https://doi.org/10.1073/pnas.1611835114, http://dx.doi.org/10.1073/pnas.1611835114 1
7. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf 3
8. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018). https://doi.org/10.1109/TPAMI.2017.2773081 3
9. Pellegrini, L., Graffieti, G., Lomonaco, V., Maltoni, D.: Latent replay for real-time continual learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 10203–10209 (2020). https://doi.org/10.1109/IROS45743.2020.9341460 1
10. Petit, G., Popescu, A., Schindler, H., Picard, D., Delezoide, B.: FeTrIL: Feature translation for exemplar-free class-incremental learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3911–3920 (2023) 1
11. Qiu, Z., Xu, Y., Meng, F., Li, H., Xu, L., Wu, Q.: Dual-consistency model inversion for non-exemplar class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24025–24035 (2024) 1, 2
12. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017). https://doi.org/10.1109/CVPR.2017.587 1
13. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. In: Advances in Neural Information Processing Systems. vol. 32 (2019) 1

14. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Advances in Neural Information Processing Systems. vol. 30, pp. 2990–2999 (2017) 1
15. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9374–9384 (2021) 1
16. Tong, R., Lu, H., Liu, Y., Gong, D.: Model inversion with layer-specific modeling and alignment for data-free continual learning. In: Advances in Neural Information Processing Systems 38 (2025) 1, 2, 3
17. van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nature Machine Intelligence **4**, 1185–1197 (2022). https://doi.org/10.1038/s42256-022-00568-3 1
18. Wu, C., Herranz, L., Liu, X., Wang, Y., van de Weijer, J., Raducanu, B.: Memory replay gans: Learning to generate new categories without forgetting. In: Advances in Neural Information Processing Systems. vol. 31 (2018) 1
19. Yin, H., Molchanov, P., Li, Z., Alvarez, J.M., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via DeepInversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8715–8724 (2020) 1, 2
20. Zhou, D.W., Wang, Q.W., Qi, Z.H., Ye, H.J., Zhan, D.C., Liu, Z.: Class-incremental learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 9851–9873 (Dec 2024). https://doi.org/10.1109/TPAMI.2024.3429383, https://doi.org/10.1109/TPAMI.2024.3429383 1